Recherche d'information et temps linguistique : une heuristique pour calculer la pertinence des expressions calendaires

Charles Teissèdre (1,2) Delphine Battistelli (3) Jean-Luc Minel (1)

(1) MoDyCo - UMR 7114 CNRS, Paris Ouest Nanterre La Défense, 200, av. de la République, 92001 Nanterre
(2) Mondeca, 3, cité Nollez, 75018 Paris
(3) STIH, Université Paris Sorbonne, 28, rue Serpente, 75006 Paris
charles.teissedre@ u-paris10.fr, delphine.battistelli@paris-sorbonne.fr, jean-luc.minel@u-paris10.fr

Résumé. A rebours de bon nombre d'applications actuelles offrant des services de recherche d'information selon des critères temporels - applications qui reposent, à y regarder de près, sur une approche consistant à filtrer les résultats en fonction de leur inclusion dans une fenêtre de temps, nous souhaitons illustrer dans cet article l'intérêt d'un service s'appuyant sur un calcul de similarité entre des expressions adverbiales calendaires. Nous décrivons une heuristique pour mesurer la pertinence d'un fragment de texte en prenant en compte la sémantique des expressions calendaires qui y sont présentes. A travers la mise en œuvre d'un système de recherche d'information, nous montrons comment il est possible de tirer profit de l'indexation d'expressions calendaires présentes dans les textes en définissant des scores de pertinence par rapport à une requête. L'objectif est de faciliter la recherche d'information en offrant la possibilité de croiser des critères de recherche thématique avec des critères temporels.

Abstract. Unlike many nowadays applications providing Information Retrieval services able to handle temporal criteria - applications which usually filter results after testing their inclusion in a time span, this paper illustrates the interest of a service based on a calculation of similarity between calendar adverbial phrases. We describe a heuristic to measure the relevance of a fragment of text by taking into account the semantics of calendar expressions. Through the implementation of an Information Retrieval system, we show how it is possible to take advantage of the indexing of calendar expressions found in texts by setting scores of relevance with respect to a query. The objective is to ease Information Retrieval by offering the possibility of crossing thematic research criteria with temporal criteria.

Mots-clés : Indexation d'informations calendaires ; Recherche d'information ; Annotation et extraction d'expressions calendaires
Keywords: Calendar information indexing ; Information Retrieval ; Annotation and extraction of calendar expressions

1 Introduction

Les systèmes de Recherche d'Information sur le Web destinés au grand public ne sont pas en mesure aujourd'hui de répondre à des requêtes exprimant des informations calendaires complexes, telles qu'on peut les exprimer en langue au travers notamment de la catégorie des adverbiaux calendaires (Klein, 1994). Dans de nombreux cas de figure, cette catégorie pourrait pourtant utilement intervenir dans le calcul de la pertinence d'un fragment de texte ou d'un document. En effet, la prise en compte de la sémantique des expressions adverbiales calendaires (soulignées ci-après) permettrait ainsi de traiter des requêtes comme celles qui suivent :

- « *Festival de musique <u>aux alentours de la mi-août</u>* »
- « *La France <u>à la fin du XVIIe siècle</u>* »
- « *Cinéma italien <u>au début des années 60</u>* »

Si les systèmes de bases de données spécialisés peuvent permettre de fournir des réponses à de telles requêtes en filtrant les réponses incluses dans une fenêtre de temps, pour autant, parce qu'elles s'appuient sur une représentation discrète, elles n'offrent pas de moyen d'ordonnancer les résultats selon des critères permettant de mesurer la pertinence relative des propriétés temporelles retournées par rapport à la requête. De tels systèmes ne permettent pas, par exemple, de jongler avec la granularité des expressions temporelles pour retenir en priorité celles qui partagent les mêmes caractéristiques que la requête ; en effet, le plus souvent, les réponses fournies dans ce cadre de recherche sont uniquement celles qui répondent favorablement au test d'inclusion dans la période recherchée.

Par delà le test d'inclusion. Une autre approche nous semble possible. Elle s'inspire du fonctionnement des moteurs de recherche, à savoir leur capacité à trier les documents par pertinence en évaluant la distance qui les rapproche ou les éloigne d'une requête. L'objectif est alors de fournir des critères pour calculer des scores de proximité entre les zones temporelles recherchées et des expressions calendaires présentes dans un texte ou un corpus de textes. La méthode que nous décrivons pour attribuer un score de pertinence temporelle permet ainsi de combiner des requêtes thématiques et des requêtes temporelles calendaires, en agrégeant et pondérant les scores de pertinence.

Afin de mettre en regard notre approche avec les travaux de recherche et les applications existants, l'article s'ouvre sur un état des lieux de la recherche d'information fondée sur des critères calendaires (section 2). Il se poursuit par la présentation de l'heuristique que nous proposons pour calculer des scores de pertinence temporelle (section 3). Enfin (section 4), nous illustrons l'intérêt de cette approche à travers une expérimentation qui prend corps dans un outil de recherche d'information destiné à montrer le type de résultats qu'il est possible d'obtenir en croisant des critères de recherche thématique avec des critères temporels.

2 Etat de l'art

2.1 Recherche d'Information et temps calendaire

Le traitement automatique de l'« information temporelle » exprimée dans les textes s'impose depuis quelques années comme un champ de recherche important auquel on associe des retombées dans le domaine de la recherche d'information (Alonso et al., 2007 ; Mestl et al., 2009) ; parmi les applications visées : les systèmes de questions/réponses, les systèmes de résumé automatique, les moteurs de recherche sur le web et, intégrés ou non à ces derniers, les systèmes visant à proposer en sortie une visualisation des informations sur une ligne du temps.

La caractérisation de l'« information temporelle » en tant que telle constitue un enjeu – au cœur des programmes d'annotation automatique - tant sur le plan descriptif (quelles sont les unités de la langue qui expriment une information temporelle ?) que sur le plan analytique (quels sont les niveaux de représentation et les stratégies calculatoires à mettre en œuvre pour appréhender la catégorie sémantique du temps ?). Dans le champ de la recherche d'information, par rapport auquel se situent précisément le plus souvent les dits

programmes d'annotation automatique, l'information temporelle est la plupart du temps rapportée à ce qui permettrait la résolution d'une tâche en particulier : celle du calcul de l'ancrage calendaire de situations (souvent appelées « événements ») décrites dans les textes. On pourra se reporter à (Battistelli, 2011) pour une présentation des enjeux descriptifs de la temporalité linguistique pour des systèmes de recherche d'information. Dans les applications citées ci-dessus, les expressions linguistiques référant explicitement à un calendrier (le calendrier grégorien par exemple) ont ainsi toujours constitué un champ d'investigation particulièrement exploré. C'est d'ailleurs dans le cadre des systèmes de questions/réponses qu'a été organisée pour la première fois en 2004 une tâche d'évaluation uniquement dévolue à la problématique de repérage puis de normalisation (i.e. de calcul en référence à une norme) de ce type d'expressions : *Time Expression Recognition and Normalization* (TERN) ; tâche plus largement à l'origine de la démarche visant à proposer une standardisation quant à l'annotation sémantique de ces expressions (cf. en particulier (Schilder et Habel, 2001 ; Pustejovsky et al., 2002 ; Ferro et. al., 2003 ; Saquete et. al., 2004 ; Ehrmann et Hagège, 2009 ; Bittar, 2010), avec en corollaire, l'élaboration de corpus annotés tels que ACE[1] et TimeBank[2] et d'un certain nombre de systèmes automatiques (cf. par exemple Mani et Wilson, 2000; Han et al., 2006; Ahn et al., 2007). Depuis peu, avec les avancées récentes sur le terrain de l'acquisition d'informations temporelles de type calendaire dans les textes dont les résultats commencent à être exploitables (UzZaman et Allen, 2010 ; Llorens et al., 2010), le champ des applications s'étend progressivement à d'autres initiatives originales, telles que la construction automatique de chronologies pour explorer et visualiser le contenu de corpus de presse (ainsi des travaux de Alonso et al, 2010 qui s'appuient sur l'outil de production de chronologies SIMILE timeline[3] ou encore ceux de Matthews et al., 2010). Parmi les quelques initiatives des moteurs de recherche sur le terrain de la recherche d'information selon des critères temporels, citons celle de Google qui permet de visualiser les résultats d'une recherche sur une chronologie, puis de filtrer les résultats sur une fenêtre de temps (il s'agit du service view:timeline). Pour autant, si ce service tire parti des expressions calendaires présentes dans les textes, il ne propose pas à proprement parler de formuler des requêtes temporelles. Du reste, seule une sous-partie des expressions calendaires rencontrées dans les textes est indexée et donne lieu à une analyse (peu ou prou, les expressions de la forme JJ MM AAAA, MM AAAA ou AAAA). Ces expressions sont en outre systématiquement réduites à une représentation atomique : ainsi, une expression telle que « *de 1815 à 1871* » n'est pas analysée comme formant une zone temporelle bornée à gauche et à droite, mais plutôt comme deux dates. On retrouve un comportement similaire dans des systèmes de gestion de connaissances structurées tel que le projet TimeSearch History[4] qui permet de croiser une recherche par mots-clés et une recherche temporelle : le champ dédié au filtre temporel permet uniquement de spécifier une année.

La difficulté du traitement des informations calendaires exprimées en langue tient dans ceci qu'elles font intervenir des opérations de régionalisation (du type avant, après, etc.), de focalisation (du type début, fin, milieu), des opérations de pointage (consistant à désigner une zone de référence sur le calendrier) et des grains calendaires variés (jour, mois, année, parties du jour, etc.). De là l'impossibilité d'organiser les expressions calendaires selon un ordre total : comment en effet ordonner d'après un unique critère des expressions aussi variées que « *avant 2009* », « *en mars 2009* », « *de mi 2009 à fin 2011* », « *aux alentours de 2009* », etc. ? De là également la tentation de simplifier le problème du traitement de ces expressions pour la recherche d'information, en les réduisant à une représentation atomique qu'il devient alors possible d'ordonner, quitte à perdre une grande partie de leur sémantique. Ce problème renvoie à celui de l'opposition entre duratif et ponctuel, identifié en Intelligence Artificielle comme un « problème de granularité » (Bettini et al., 2000 ; Bechet et al., 2000), dont il est pourtant possible de sortir en considérant qu'il s'agit essentiellement d'une question d'échelle.

En dépit des quelques initiatives mentionnées, les expressions calendaires dans les textes sont encore très largement aujourd'hui traitées par les moteurs de recherche grand public comme des mots-clés dont la sémantique n'est pas ou peu exploitée. Ainsi, une recherche par mots-clés sur un intervalle de temps (mettons « *de 1750 à 1800* ») ne ramène que des résultats où les termes mêmes de la recherche apparaissent (on pourra ainsi trouver des expressions telles que « *en 1750* » ou « *en 1800* », mais pas des expressions telles que « *peu après 1763* » ou « *de 1755 à 1799* »).

[1] Cf. les corpus LDC2005T07 et LDC2006T06 du catalogue LDC (http://www.ldc.upenn.edu).

[2] Cf. le corpus LDC2006T08 du catalogue LDC.

[3] SIMILE Timeline toolkit: http://simile.mit.edu/timeline/

[4] http://www.timesearch.info/

2.2 Inclusion vs. similarité temporelle

Pour comprendre l'intérêt d'un principe de pertinence temporelle que nous souhaitons appréhender, il faut souligner les limites des approches réduisant l'interrogation temporelle à l'inclusion dans une fenêtre de temps - approche retenue dans les systèmes de gestion de BDD ou dans les moteurs de recherche qui proposent d'indexer des données temporelles. En filtrant la recherche par l'inclusion, le risque est (1) de ne pas renvoyer de résultat, (2) de ne pas pouvoir ordonner les résultats par pertinence sous l'angle des propriétés temporelles, et (3) d'avoir des difficultés à jongler entre des granularités ou échelles de temps différentes. Ces limites tiennent à la méthodologie retenue, qui relève, au fond, d'une approche booléenne (inclusion vs. non inclusion). L'intérêt du calcul d'un score de similarité est d'échapper à cette approche restrictive, en permettant d'évaluer la distance/proximité entre différentes caractéristiques des expressions calendaires. Si les relations entre intervalles de temps telles que les a décrites (Allen, 1983) permettent de comparer des expressions calendaires et d'obtenir des résultats booléens (en testant l'inclusion, le recouvrement, l'intersection, etc.), elles ne permettent pas à elles seules de les hiérarchiser par pertinence, ce qui constitue le cœur de notre approche. Pour cela il faut recourir à des mesures telles que la distance temporelle entre intervalles de temps, le taux de recouvrement, les rapports de proportions entre expressions, etc.

(Le Parc-Lacayrelle et al., 2007) proposent une méthode de calcul de la pertinence temporelle par rapport aux centroïdes des intervalles de temps. La méthode décrite consiste à calculer la pertinence relative entre la requête et la partie de l'index qui entre en intersection avec la requête. Pour deux expressions incluses dans la période définie par la requête (par exemple, les expressions « *en 1804* » et « *en 1859* » pour une requête telle que « *au XIX*e *siècle* »), on privilégie celle qui est la plus proche du centre (« *en 1859* »). Ces travaux ont ceci d'intéressant qu'ils introduisent la notion de pertinence en proposant une méthode d'ordonnancement de fragments de documents (qui renvoie aux problématiques de « ranking » ou de « scoring » familières aux moteurs de recherche). Ils se limitent toutefois à l'intersection entre une requête et des expressions calendaires présentes dans les textes (soit encore à l'inclusion dans une fenêtre de temps, bien que l'inclusion soit ici entendue en un sens moins restrictif). Ceci a pour effet, pour une requête telle que « *le 12 août 1988* », d'exclure des résultats des expressions telles que « *dans la nuit du 10 au 11 août 1988* » qui sont pourtant susceptibles de présenter un intérêt.

2.3 L'annotation des expressions calendaires

Dans le cadre du traitement automatique des informations calendaires dans les textes, il est nécessaire que les systèmes d'annotation soient en mesure de traiter des expressions parfois très complexes tout en proposant une représentation formelle manipulable par les machines et suffisamment riche pour couvrir au mieux la manière dont la langue exprime une référence au calendrier : le modèle d'annotation a donc ici toute son importance.

Quelles que soient les difficultés rencontrées par les systèmes, en particulier quant au calcul de la valeur d'une expression relative - déictique ou anaphorique - (Cailliau et al., 2008 ; Wang et Zang, 2008 ; Mazur et Dale, 2008), il reste qu'aucune démarche n'a à notre connaissance pris pour objet l'analyse des *relations* entre ces expressions en tirant parti, non seulement de leurs valeurs, mais aussi des unités linguistiques à proprement parler dans lesquelles elles entrent le plus souvent, à savoir des unités adverbiales. A dire vrai, ce ne sont d'ailleurs pas des expressions temporelles adverbiales qui font l'objet d'une annotation (si l'on s'en tient en particulier aux deux projets d'annotation majeurs à l'heure actuelle que sont TIMEX2 (Ferro et al., 2003) et TIMEX3 (Pustejovsky et al., 2003, 2005), mais uniquement leur référence au système calendaire à proprement parler, la signification des locutions prépositionnelles étant vouée à être traitée à part (via une balise nommée SIGNAL). D'autres approches, comme (Aunargue et al., 2001), (Battistelli et al., 2008), (Teissèdre et al., 2010) tiennent compte, elles, des prépositions et de l'ensemble des éléments qui interviennent dans la composition des adverbiaux temporels – éléments qui s'avèrent très utiles pour l'analyse sémantique et l'indexation d'expressions telles que « *au début du XXe siècle* », « *vers le milieu des années 1950* », « *3 mois avant la fin de l'année* ».

Le modèle d'annotation proposé dans (Battistelli et al., 2008) permet ainsi un traitement fin de la granularité, en décrivant la composition des expressions calendaires et la façon dont elles imbriquent, par-dessus une référence à un système calendaire (autrement dit, la base calendaire), des opérations de pointage (pointage déictique, anaphorique ou absolu), de focalisation (début, milieu, fin), de déplacement (ex : « *deux jours avant* ») et de régionalisation (avant, après, pendant, jusqu'à, etc.).

Enfin, on peut mentionner des travaux qui portent sur les expressions itératives (Teissèdre et al., 2010), qui requièrent également des traitements particuliers, faisant notamment appel à des outils de raisonnement pour passer de propriétés temporelles définies en intension (ex: « *tous les lundis* ») à une représentation en extension (ex : le lundi 15 mars, le lundi 22 mars, etc.). Ces traitements permettent d'appréhender des expressions complexes comme les dates et horaires d'ouverture (ex : « *ouverture du lundi au vendredi, de 9h à 19h.* »).

S'appuyant sur les informations issues des moteurs d'annotation, le système d'indexation et de recherche que nous avons développé met en regard une requête et des expressions calendaires présentes dans les documents indexés, afin de faire ressortir les parties les plus pertinentes de ces documents. La modélisation retenue par les systèmes d'annotation influe donc lourdement en aval sur la capacité à comparer entre elles les caractéristiques de différentes expressions calendaires.

3 Vers une heurististique de calcul de la pertinence temporelle

3.1 L'intérêt d'évaluer la pertinence temporelle

Afin d'illustrer l'intérêt du calcul de la pertinence temporelle, la figure 1 montre la façon dont le système que nous avons développé hiérarchise, pour deux requêtes différentes, un même ensemble d'expressions calendaires. Les résultats sont présentés à la manière d'un nuage de tags où les éléments les plus pertinents sont surlignés graphiquement : la pertinence d'un résultat joue sur la taille de la police et le niveau de gris[5].

Requête 1 : « Dans les années 1930 »	Requête 2 : « en 1931 »
en 1929	en 1929
Au cours des années 1930	Au cours des années 1930
entre 1930 et 1934	entre 1930 et 1934
en 1932	en 1932
À partir de 1936	À partir de 1936
en 1937	en 1937
le 17 juin 1939	le 17 juin 1939
Fin 1940	Fin 1940

Fig. 1 : nuages d'expressions calendaires proposées pour deux requêtes

La figure 1 montre que la hiérarchie des résultats varie en fonction des caractéristiques de la requête : elle dépend de plusieurs critères, qui tiennent compte de l'inclusion, certes, mais aussi de la distance temporelle et des rapports de proportion (ou granularité) entre la requête et les expressions calendaires indexées. L'outil de navigation présenté dans la dernière section s'appuie sur cet algorithme d'ordonnancement pour présenter les portions de textes les plus pertinentes pour une requête comportant un critère temporel. La figure permet également d'illustrer l'intérêt de l'approche par rapport aux systèmes qui filtrent les résultats sur le seul critère d'inclusion : la requête 2 présente une étendue temporelle restreinte, dans laquelle aucun des résultats présentés n'est inclus. Pour autant, le système est tout de même en mesure de fournir des résultats classés par pertinence, même s'ils ne sont pas inclus dans la période définie par la requête de l'utilisateur.

3.2 Heuristique de pertinence temporelle

Les critères entrant dans le calcul de la pertinence temporelle sont : (i) le score brut d'intersection, (ii) la mesure de la distance entre la requête et les expressions calendaires apparaissant dans les documents, (iii) la mesure du rapport de proportion (iv) la mesure de la distance pondérée par la granularité et (v) la mesure du

[5] D'autres choix visuels sont à l'étude, afin de permettre de jongler entre différents types d'ordonnancement tout en laissant apparaître le degré de pertinence des expressions calendaires par rapport à une requête (tri par ordre de pertinence, tri par ordre d'apparition dans un texte, tri selon l'ordre temporel, etc.).

taux de recouvrement. Chacun de ces scores produit un résultat entre 0 et 1, où 1 est le meilleur score possible. La logique de ces scores est illustrée par différentes figures où les expressions calendaires (EC) dont on évalue la pertinence relative sont représentées au dessus de l'axe du temps : la dénomination « EC Requête » y désigne la zone temporelle définie par la requête et « EC Rép. » les zones temporelles définies par les expressions formant un ensemble possible de réponses.

- Score brut d'intersection

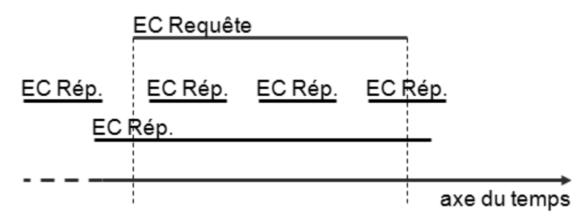


Fig. 2 : intersection

Ce score recouvre une sous partie des relations décrites dans (Allen, 1983) : la relation « during » (inclusion), la relation « overlaps » (recoupement), la relation « meets » (adjacence) et leur relation inverse. Les scores attribués diffèrent pour l'inclusion complète, partielle, ou inverse (lorsque la requête est incluse dans une expression d'étendue plus vaste) ; le score est nul pour une expression qui n'a pas d'intersection avec la requête. Ce score permet de catégoriser les documents à ordonner en cinq classes (intervalles inclus, incluant, en intersection, concomitant ou exclus). Ce critère est insuffisant à lui seul, parce que l'intersection n'est pas toujours synonyme de plus grande pertinence : ainsi, pour une requête temporelle telle que « *en 1953* », l'expression « *durant le XXe siècle* » obtient un meilleur score d'intersection que l'expression « *en 1954* », qui est pourtant plus proche de la requête du point de vue de la granularité. La catégorisation obtenue à l'aide du score d'intersection est utile pour décider ensuite, parmi les autres critères de mesure de la pertinence ceux qui seront calculés (score de distance ou score de recouvrement), mais aussi ceux des scores intermédiaires qu'il faudra privilégier au détriment des autres.

- Score de distance

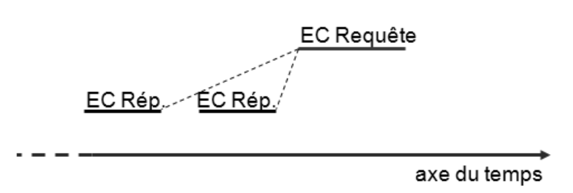


Fig. 3 : distance

La distance correspond à l'étendue de temps séparant deux expressions calendaires. Ce score est calculé pour les expressions dont l'intersection avec la requête est nulle. Pour celles où il y a intersection, le pendant du score de distance est le taux de recouvrement. La mesure de la distance porte sur la borne droite pour les expressions s'achevant avant la période désignée par la requête et sur la borne gauche pour les expressions débutant après.

- Score de proportion

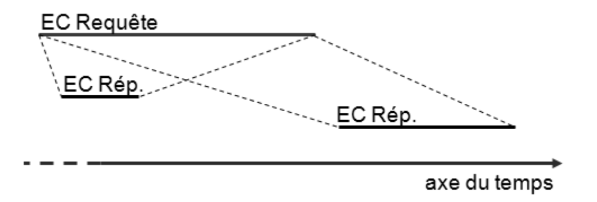


Fig. 4 : proportion

La mesure de la proportion relative permet d'évaluer la proximité des résultats avec la requête du point de vue de la granularité. Pour une requête telle que « *le 15 juillet 2007* », il semble en effet plus pertinent d'avoir en tête de liste une date comme « *le 16 juillet 2007* » plutôt que « *en 2007* » ou encore « *au XXIe*

siècle ». On postule donc ici qu'un utilisateur sera vraisemblablement plus intéressé par une réponse qui n'est pas nécessairement incluse dans la période définie par sa requête, mais qui en est proche sémantiquement, au sens où la nature de la requête détermine les caractéristiques attendues dans les résultats.

- Score de distance pondéré par la granularité

Ce score vise à pondérer la mesure de l'étendue de temps séparant les deux expressions comparées en la faisant dépendre de la granularité de la requête : la distance entre deux expressions telles que « *en 1840* » et « *en 1860* » est évaluée comme étant plus faible qu'entre les expressions « *le 11 juin 1840* » et « *le 20 mars 1860* », car la granularité des premières est plus étendue.

- Score de recouvrement

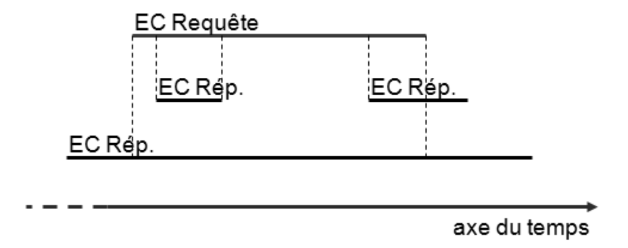


Fig. 5 : score de recouvrement

Ce score reprend ici la proposition de modèle de calcul de la pertinence formulée par (Le Parc et al., 07), à ceci près que l'on fait varier le centroïde d'une expression en fonction de la nature des opérations de régionalisation ou de focalisation qui interviennent dans la composition des expressions comparées. On privilégie ainsi la distance par rapport à la borne du début pour la plupart des expressions : ceci permet, pour une requête telle que « *dans les années 80* », de faire remonter dans la hiérarchie des résultats « *en 1981* » plutôt que « *en 1985* » ; ce qui revient à dire qu'on privilégie l'ordonnancement par rapport à l'axe du temps (de gauche à droite). Pour les expressions présentant une régionalisation de type « avant » ou « jusqu'à » (expressions définissant des intervalles infinis à gauche), la distance des intervalles comparés dépend de la borne droite. On considère en revanche le centroïde pour les expressions présentant une focalisation de type « milieu » (« *à la mi-mai* »). Ce score n'a de sens et n'est calculé qu'en cas d'intersection entre la requête et l'expression testée.

Le cas des intervalles infinis à gauche ou à droite. Les intervalles infinis à gauche ou à droite (ex. : « *depuis la fin du XVIe siècle* », « *jusqu'au 13 mars 2003*) demandent un traitement spécifique pour le calcul des proportions relatives et du taux de recouvrement, parce que leur durée est infinie. Ainsi, pour le calcul du score de proportion, on ne retient que l'étendue de la borne principale (« *1950* » pour l'expression « *jusqu'en 1950* »). On suppose donc que l'étendue de cette borne détermine la proportion des expressions susceptibles d'intéresser l'utilisateur : une requête telle que « *avant le XVIe siècle* » renverra en tête de liste des résultats de même granularité. Pour le calcul de l'intersection, on attribue une valeur minime, mais non nulle, qui place tous les résultats qui sont en intersection avec la requête au même niveau.

Combiner les modèles de pertinence. La difficulté pour établir un score global de pertinence temporelle, c'est que, pris isolément, chacun de ces critères peut produire un ordre ou un classement différent des autres : il s'agit donc de les combiner afin de produire des résultats cohérents et pertinents. Notre proposition d'heuristique consiste à agréger les scores intermédiaires, avec un facteur d'augmentation ou de réduction des scores.

Score de pertinence temporelle = K × Score d'intersection +
L × Score de distance +
M × Score de proportion +
N × Score de distance pondéré +
O × Score de recouvrement

Les facteurs de maximisation d'un score et le calcul même des scores sont interdépendants. Ils peuvent ainsi varier selon que les résultats sont ou non inclus dans la période déterminée par la requête. En cas de non inclusion, plus la distance est grande, moins les rapports de proportions importent :

Score de Proportion = Score Proportion × Score Distance

De la même façon, si le score de proportion est très faible, on minimise le score de distance et le score de recoupement. A ce stade les facteurs intervenant dans le calcul de pertinence ont été déterminés empiriquement. Nous travaillons actuellement à une formalisation plus systématique, à la fois du modèle de pertinence et de la prise en compte des pôles des intervalles (début, milieu, fin), qui fera l'objet d'un prochain article (Battistelli et al., soumis).

4 Un outil de navigation temporelle dans les textes

L'outil de navigation textuelle que l'on décrit dans cette section permet à l'utilisateur/lecteur d'effectuer des recherches thématiques et temporelles dans un texte ou un ensemble de textes, pour filtrer ou surligner les passages les plus pertinents. Les fonctionnalités de recherche s'appuient sur les techniques d'indexation auxquelles ont communément recours les moteurs de recherche. De ce point de vue, l'outil développé peut tout aussi bien être présenté comme un prototype de moteur de recherche en mesure de traiter des requêtes temporelles.

Le modèle retenu pour établir la pertinence par rapport aux mots-clés est ici plutôt simpliste et repose sur les outils standards proposés dans la suite Lucene. Cette expérimentation a en effet pour fonction d'illustrer le type de résultats qu'il est possible d'obtenir en croisant une requête thématique (recherche par mots-clés) et une requête temporelle. Ceci implique, sur un plan technique, de combiner les scores de pertinence obtenue par ces différents modèles - temporel et thématique -, afin d'obtenir une liste de résultats hiérarchisés. Naturellement, les deux modèles de calcul de la pertinence entrent en contradiction, ce qui oblige à trouver un juste équilibre entre eux. Sur ce plan toutefois, il s'agit d'un problème classique que tout moteur de recherche doit gérer, puisqu'il leur faut bien croiser les modèles de pertinence par rapport aux mots-clés (« word similarity »), par rapport à la popularité des sites Web (« popularity »), par rapport à la qualité du site (« trust »), pour ne reprendre que quelques uns des modèles d'ordonnancement les plus connus, qui chacun produit un ordre différent des autres.

Le démonstrateur, un service Web accessible en ligne[6], permet à l'utilisateur de charger un texte ou un corpus de textes. Un premier pré-traitement segmente les textes en paragraphes et en phrases, dont les expressions calendaires sont alors annotées, puis indexées. Pour cette expérimentation, les expressions relatives, qui demandent un calcul particulier (les déictiques et anaphoriques tels que « *demain* », « *deux jours plus tard* »), ne sont pas traitées. Seules les expressions absolues, qui peuvent être disposées sur le calendrier sans recourir à des outils de résolutions des anaphores, sont à ce jour indexées.

4.1 Pré-traitements : des expressions calendaires aux intervalles calendaires

Les ressources pour l'annotation des expressions calendaires utilisées dans notre système consistent en un ensemble de transducteurs Unitex (Paumier, 2000) présentés dans (Teissèdre et al., 2010). Les expressions annotées par ces ressources sont ensuite transformées en intervalles calendaires. Ce passage du modèle linguistique ou symbolique vers le modèle calendaire constitue une problématique à part entière. Elle est prise en charge par un module décrit dans (Battistelli et al., 2011). Retenons donc ici que les expressions calendaires annotées sont transformées en un ou plusieurs intervalles (pour les itératifs en particulier, comme « *tous les jours sauf le dimanche* »), dont les bornes peuvent être transformées au format ISO-8601.

Exemples de transformation vers des intervalles calendaires

- « *de 1830 à 1940* » : [1830, 1940]
- « *XVIe siècle* » : [15**, 15**][7]

[6] http://client1.mondeca.com/TemporalQueryModule/

[7] La notation YY** renvoie à des siècles et la notation YYY* à des décennies. Les intervalles peuvent également s'écrire de la façon suivante [début(date1), fin(date2)], dans la mesure où les bornes des intervalles (date1 et date2) forment elles-mêmes des intervalles ayant une durée.

- « *fin du mois de juin 2010* » : [2010−06−20, 2010−06−30]
- « *depuis le milieu des années 60* » : [1965, + ∞[

Il faut souligner que l'exactitude (du reste impossible à établir) dans la transformation d'une représentation symbolique des expressions calendaires vers une représentation discrète (les intervalles calendaires) ne revêt pas une importance majeure, dans la mesure où la recherche repose sur des mesures de similarité : ainsi, savoir si « *le 18 juin 2010* » est inclus ou non dans la période définie par « *fin juin 2010* » n'est pas crucial, puisque que la « distance sémantique » entre les deux sera de toute façon évaluée comme étant faible. Dit autrement, une expression comme « *le 18 juin 2010* » sera évaluée comme étant moins prototypique (ou similaire) que « *le 30 juin 2010* » au regard des expressions attendues pour une requête portant sur la « *fin juin 2010* », elle sera toutefois considérée comme étant plus pertinente qu'une expression comme « *le 10 juin 2010* ».

4.2 Indexation avec Lucene

L'indexation repose sur l'API de Lucene[8] et les analyseurs proposés dans la librairie pour le français et l'anglais. Le paradigme standard pour le traitement des mots-clés (qui exclut les « mots-vides » ou « stop-words ») est donc repris tel quel, dans la mesure où il ne s'agit pas directement de traiter la question des recherches thématiques, mais plutôt d'illustrer les possibilités ouvertes par le traitement des requêtes calendaires.

A ce stade des développements, les objets indexés sont des fragments de documents, en l'occurrence, des phrases, afin de disposer du segment textuel contenant l'expression temporelle qui nous intéresse, mais aussi afin de s'assurer que la distance entre l'expression temporelle et les mots-clés, lorsqu'ils sont retrouvés dans le texte, ne soit pas trop importante. A ces fragments sont associés les termes du paragraphe, l'url de la page Web ou le titre du document et les différentes expressions calendaires présentes. Lorsqu'un fragment de document contient plusieurs expressions calendaires, il est indexé plusieurs fois. Pour une requête thématique, on privilégie ainsi les mots-clés présents dans la phrase, mais on tient compte également de ceux présents dans le paragraphe et l'url ou le titre. Il s'agit là d'une simplification rudimentaire du problème complexe, du point de vue de la linguistique textuelle, de la portée des expressions calendaires, c'est-à-dire de la façon dont elles prennent part à la structuration du discours (Van Reamdonck, 2001 ; Charolles et Vigier, 2005 ; Bilhaut et al., 2003).

Mettre en regard une requête temporelle et des expressions calendaires pour évaluer leur pertinence ne peut pas se faire d'emblée au niveau de l'indexation, car la mesure de la pertinence relative des résultats par rapport à la requête nécessite une comparaison deux à deux des expressions calendaires. Pour éviter des traitements coûteux qui demanderaient de balayer tout l'index pour comparer les informations temporelles stockées et la requête, le parti pris est de réduire les expressions calendaires, lors de l'indexation, à des éléments atomiques (ponctuels) sur l'axe du temps. On réduit ainsi provisoirement les expressions calendaires à une ancre calendaire sans étendue, en ne retenant des expressions que leur « point focal » qui correspond à l'ancre calendaire évaluée comme la plus pertinente. Ce point correspond le plus souvent à la borne gauche de l'intervalle temporel décrit par une expression calendaire, ce qui revient à dire qu'on privilégie l'ordre lié au sens de l'écoulement du temps : ainsi pour une requête du type « *au XXe siècle* », on ordonnera par exemple une série de résultats de la façon suivante « *en 1903* », « *de 1910 à 1912* », « *vers 1920* », notamment parce que la distance des ces expressions par rapport à la borne de gauche de la requête va croissant. Toutefois, pour des expressions formant un intervalle infini à gauche (ex : « *jusqu'en 2007* ») ou une focalisation de type « fin » (ex : « *fin août 1988* »), le point d'ancrage est la borne droite. Ce point correspond en revanche au centre pour les expressions présentant une focalisation de type « milieu » (« *mi-août 1993* »). Ainsi, par exemple, l'expression « *au XVIe siècle* » est réduite pour l'indexation à la date suivante au format ISO-8601, 1500-01-01T00:00:00, qui correspond au début de la borne gauche de l'intervalle [15**, 15**]. Le point d'ancrage de l'expression « *au milieu des années 50* » correspond lui au centroïde de l'intervalle calendaire, à savoir la date suivante : 1955-01-01T00:00:00.

Le processus de comparaison et d'évaluation du score de pertinence n'intervient qu'après filtrage des résultats. Le requêtage de l'index se fait ainsi en deux étapes : (1) recherche dans l'index des K plus proches voisins de la requête (par rapport au point focal de l'intervalle calendaire) ; (2) ordonnancement des

8 http://lucene.apache.org/

résultats ainsi filtrés en fonction de leur pertinence relativement à la requête. La première étape du requêtage consiste à récupérer dans l'index des documents les K plus proches voisins de la requête, où K correspond au nombre de résultats que l'on souhaite présenter (soit les K résultats a priori les plus intéressants). L'étape suivante consiste à évaluer la pertinence des résultats issus de l'étape de filtrage, qui permet d'obtenir un ordre total des expressions calendaires sous l'angle de la pertinence relative des documents face à une requête.

4.3 Croiser une requête thématique et une requête temporelle

Les requêtes soumises au système par les utilisateurs sont annotées par le même module d'annotation que celui utilisé pour annoter les expressions calendaires dans les documents indexés. L'analyse de la requête sépare un ensemble de mots-clés (la requête thématique) et une requête temporelle, exprimée en langage naturel. Lors du parcours de l'index, les cinquante plus proches voisins de la requête sur l'axe du temps sont rassemblés, selon la méthode décrite précédemment ; cet ensemble ainsi filtré est ensuite ordonné par pertinence.

Pour cette première expérimentation, 7782 articles en français provenant de Wikipedia et relatifs à l'histoire de France ont été annotés et indexés. Le corpus contient près de 180 000 expressions calendaires « absolues ». La figure 6 est une copie d'écran des résultats renvoyés pour la requête suivante « *Jules Ferry à la fin des années 1880* ». Les premiers résultats présentés sont ceux dont l'expression calendaire est considérée comme la plus proche sémantiquement de la requête.

Jules Ferry à la fin des années 1880 | Rechercher | page d'accueil

Jules Ferry - Wikipédia

- Lâché par les radicaux, **Jules Ferry** échoue lors de la désignation du Président de la République **le 3 décembre 1887**.
- Le caractère économique (véritable source du débat), occupe l'extrême majorité des propos de **Jules Ferry**, **le 28 juillet 1885** :
- On retrouve notamment l'empreinte de **Jules Ferry** en Tunisie dont il obtient le protectorat **le 12 mai 1881** par le traité du Bardo, à Madagascar, il lance l'explorateur Pierre Savorgnan de Brazza à la conquête du Congo (1879).
- **Jules Ferry**, candidat républicain en face de Louis-Joseph Buffet, chef de file de la majorité départementale est élu **le 8 octobre 1871** conseiller général du Thillot.

Georges Clemenceau - Wikipédia

- Son combat contre **Jules Ferry** aboutit **le 30 mars 1885** à la démission de ce dernier sur l'affaire du Tonkin.
- **En février 1883**, **Jules Ferry** forme son deuxième cabinet, appuyé sur une coalition centriste (Union républicaine et « Gauche républicaine », groupe de centre-droit en réalité).

Fig. 6 : copie d'écran de l'outil de navigation temporelle dans un corpus

Les résultats présentés à l'écran (un ensemble de phrases) sont regroupés par textes source et, pour chacun des textes, par pertinence. La figure 6 montre que la sémantique de la requête temporelle est correctement interprétée : en effet, dans les résultats présentés, on trouve en tête de liste des résultats proches de la borne droite de l'intervalle correspondant à l'expression « *à la fin des années 1880* ». Les résultats sont en outre ordonnés du plus prototypique au moins pertinent par rapport à la période définie par la requête.

5 Perspectives

L'expérience montre que la méthodologie proposée pour calculer des scores de pertinence temporelle permet de traiter des requêtes temporelles et de faire remonter des résultats qui tiennent compte des caractéristiques temporelles de la requête. Elle illustre également la façon dont il est possible de croiser une requête thématique et une requête temporelle pour explorer un texte ou un corpus.

Nous commençons à travailler sur la spécification de protocoles pour permettre d'évaluer le système lui-même. Les prochains développements viseront également à proposer un modèle de pertinence pour des documents en entier (des sites Web, plutôt que des phrases), afin de disposer d'un outil de recherche d'informations au sein de corpus étendus sur le Web. En effet, l'algorithme de calcul de la pertinence temporelle peut servir tout aussi bien à la navigation au sein d'un texte ou d'un corpus de textes (l'application que l'on présente dans le cadre de cet article) qu'à la recherche d'informations sur le Web, dans la mesure où les techniques utilisées sont semblables, même s'il faut encore parvenir à établir un

modèle de pertinence pour un document. Ces travaux visent ainsi, à terme, à montrer qu'il est possible de combler le vide qu'il y a aujourd'hui entre la recherche d'informations sur le Web et la consultation d'un document proposé dans les résultats d'une recherche, en permettant à l'utilisateur de fouiller le texte d'un site Web aisément, avant d'y accéder.

Remerciements

Ce projet est partiellement financé par l'ANR (Contint) RMM2.

Références

AHN D., VAN RANTWIJK J., DE RIJKE M. (2007). A cascaded machine learning approach to interpreting temporal expressions. Actes de *NAACL-HLT'07* Rochester, NY, USA, April.

ALLEN J. F. (1983). Maintaining knowledge about temporal intervals. Actes de *ACM 26, no. 11* (November). 832-843.

ALONSO O., GERTZ M., BAEZA-YATES R. (2007). On the value of temporal information in information retrieval. Actes de *ACM SIGIR Forum 41, no. 2* (December). 35-41.

ALONSO O., BERBERICH K., BEDATHUR S., WEIKUM G. (2010). Time-Based Exploration of News Archives. In *HCIR 2010*. New Brunswick. 12-15.

AUNARGUE M., BRAS M., VIEU L., ASHER N. (2001). The syntax and semantics of locating adverbials. *Cahiers de Grammaire*, 26, 11–35.

BATTISTELLI D., CORI M., MINEL J.-L., TEISSEDRE C. (soumis). Querying calendar references in texts. 8 p.

BATTISTELLI D., CORI M., MINEL J.-L., TEISSEDRE C. (2011). Semantics of Calendar Adverbials for Information Retrieval. *ISMIS 2011* (LNCS), Warsaw, June 28-30 2011, 9 p.

BATTISTELLI D. (2011). Linguistique et recherche d'information : la problématique du temps. *Hermès Sciences*, 249 pages, coll. Traitement de l'Information, avril 2011.

BATTISTELLI D., COUTO J., MINEL J.-L., SCHWER, S. (2008). Representing and Visualizing calendar expressions in texts. Actes de *STEP'08*, Venice.

BECHET G., CLERIN-DEBARD F., ENJALBERT P. (2000). A qualitative Model for Time Granularity. *Computational Intelligence*, Vol. 16 (2), 137-175.

BETTINI C., JAJODIA S., WANG S. (2000). Time granularities in Databases, Datamining, and Temporal Reasoning. *Springer* (Eds), 2000, XI, 230 p.

BILHAUT F., HO-DAC L.M., BORILLO A., CHARNOIS T., ENJALBERT P., LE DRAOULEC A., MATHET Y., MIGUET H., PÉRY-WOODLEY M.-P., SARDA L. (2003). Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique. Actes de *TALN 2003*, 315-320.

BITTAR A. (2010). Building a TimeBank for French: A Reference Corpus Annotated According to the ISO-TimeML Standard. *Thèse de doctorat*, Université Paris Diderot (Paris 7).

CAILLIAU F., GIRAUDEL A., ARNULPHY B. (2009). Tracking Out-of-date Newspaper Articles. Advances in Computational Linguistics, *Research in Computing Science 41*, 2009, 277-288.

CHAROLLES M., VIGIER D. (2005). Les adverbiaux en position préverbale : portée cadrative et organisation des discours. *Langue Française* vol 2, no. 148, 9-30.

EHRMANN M., HAGÈGE C. (2009). Proposition de caractérisation et de typage des expressions temporelles en contexte. Actes de *TALN'09*, Senlis, France

FERRO L., GERBER L., MANI I., SUNDHEIM B., WILSON G. (2003). TIDES Standard for the Annotation of Temporal Expressions, http://www.mitre.org/work/tech\-papers/tech-papers-04/ferro-tides/.

HAN B., GATES D., LEVIN L. (2006). From language to time: A temporal expression anchorer. Actes *TIME'06,* IEEE Computer Society, June, 196–203.

KLEIN W. (1994). Time in Language. London, *Routledge*.

LE PARC-LACAYRELLE A., GAIO M., SALLABERRY C. (2007). La composante temps dans l'information géographique textuelle, *Document numérique* 2/2007 (Vol. 10), 129-148.

LLORENS H., SAQUETE E., NAVARRO B. (2010). TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. Actes de *5th International Workshop on Semantic Evaluation, ACL 2010*, 284-291.

MANI I., WILSON G. (2000). Robust temporal processing of news. Actes de 38th ACL, 69–76.

MATTHEWS M., TOLCHINSKY P., MIKA P., BLANCO R., ZARAGOZA, H. (2010). Searching through time in the New York Times Categories and Subject Descriptors. Actes de *HCIR 2010* - Challenge Report, 41-44.

MAZUR P., DALE R. (2008). What's the Date? High Accuracy Interpretation of Weekday Names. Actes de *22nd International Conference on Computational Linguistics*. Manchester. 553–560.

MESTL T., CERRATO O., ØLNES J. , MYRSETH P., GUSTAVSEN I.-M. (2009). Time Challenges – Challenging Times for Future Information Search. *D-Lib Magazine*, Volume 15 Number 5/6, May/June 2009.

PAUMIER S. (2002). Manuel d'utilisation du logiciel Unitex. IGM, Université de Marne-La-Vallée.

PUSTEJOVSKY J., CASTANO J., INGRIA R., SAURÍ R., GAIZAUSKAS R., SETZER A., KATZ G. (2003). TimeML: Robust specification of event and temporal expressions in text. Actes de *IWCS-5, Fifth International Workshop on Computational Semantics*.

SAQUETE E., P. MARTÍNEZ-BARCO, R. MUÑOZ, J.L. VICEDO (2004). Splitting Complex Temporal Questions for Question Answering systems. Association for Computational Linguistics (ACL) Barcelona, SPAIN. July 2004

SCHILDER F., HABEL C. (2001). From Temporal Expres-sions to Temporal Information: Semantic Tagging of News Messages. Actes de *ACL'01, Workshop on temporal and spatial information processing*, 65 -72.

TEISSEDRE C., BATTISTELLI D., MINEL J.-L. (2010). Resources for Calendar Expressions Semantic Tagging and Temporal Navigation through Texts, Actes de *LREC 2010*, 3572-3577.

TEISSEDRE C., BATTISTELLI D., MINEL J.-L. (2010). Du texte au portail sémantique : cas d'utilisation lié à des données temporelles. In Actes d'*IC'2010*, 209-220.

UZZAMAN N., ALLEN J. F. (2010). TRIPS and TRIOS System for TempEval-2: Extracting Temporal Information from Text. Actes du *5th International Workshop on Semantic Evaluation, ACL 2010*. 276-283.

VAN RAEMDONCK D. (2001). Est-il pertinent de parler d'une classe d'adverbes de temps ? Actes de *CLAC* 7.

WANG R., ZHANG Y. (2008). Recognizing Textual Entailment with Temporal Expressions in Natural Language Texts. Actes de *2008 IEEE International Workshop on Semantic Computing and Applications* (IWSCA '08).